

Quelle langue pour la RI ?

Deuxième partie

J. Savoy, Université de Neuchâtel

<http://www.clef-campaign.org>

<http://research.nii.ac.jp/ntcir/>

<http://trec.nist.gov> (TREC-3 to TREC-12)

© copyright J. SAVOY

EARIA'06 - 1 -

Le défi

"Given a query in any medium and any language, select relevant items from a multilingual multimedia collection which can be in any medium and any language, and present them in the style or order most likely to be useful to the querier, with identical or near identical objects in different media or languages appropriately identified."

[D. Oard & D. Hull, AAAI Symposium on Cross-Language IR, Spring 1997, Stanford]

EARIA'06 - 2 -

Plan

- **Introduction**
- RI monolingue
- Problèmes de traduction
- Stratégies de traduction (RI bilingue)
- RI multilingue

EARIA'06 - 3 -

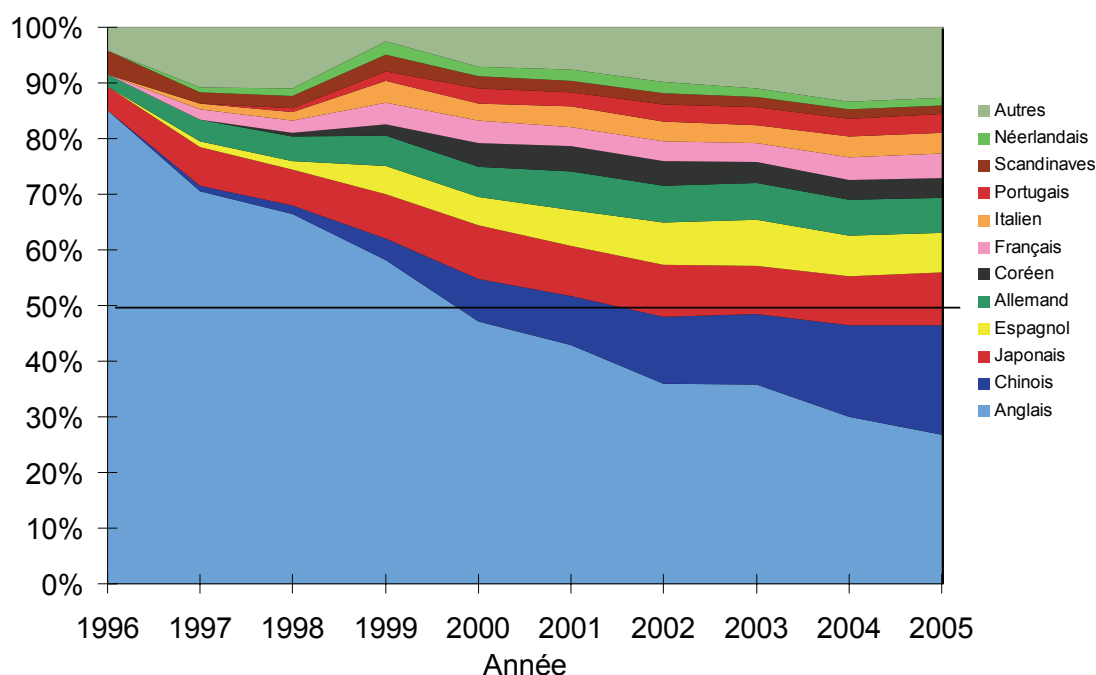
Introduction

- Quelques faits (www.ethnologue.com)
 - 6 800 langues dans le monde, dont
 - 2 197 en Asie
 - 2 092 en Afrique
 - 1 310 dans le Pacifique
 - 1 002 en Amérique
 - 230 en Europe.
 - 600 d'entre elles sont écrites
 - 80 % de la population mondiale parle 75 langues différentes
 - 40 % de la population mondiale parle 8 langues différentes
 - 75 langues sont parlées par plus de 10 M de personnes
 - 20 langues sont parlées par plus de 50 M de personnes
 - 8 langues sont parlées par plus de 100 M de personnes.

EARIA'06 - 4 -

Introduction

Pourcentage d'utilisateurs d'Internet selon la langue



EARIA'06 - 5 -

Introduction

- Bilingue et multilingue
 - Plusieurs pays sont bilingues ou multilingues (Canada (2), Singapour (2), Inde (21), UE (20))
 - Langues officielles dans UE : allemand, anglais, danois, espagnol, estonien, finnois, français, grec, hongrois, italien, letton, lithuanien, maltais, néerlandais, polonais, portugais, slovaque, slovène, suédois, tchèque, (bulgare, roumain).
Autres langues: basque, catalan, écossais, gaélique, gallois, russe.
 - Langues de travail dans UE : allemand, anglais, français;
Et à l'ONU: anglais, arabe, chinois, espagnol, français, russe.
 - Décisions judiciaires écrites dans plusieurs langues (Canada)
 - Organisations / entreprises: FIFA, WTO, UBS, Nestlé,
...

EARIA'06 - 6 -

Introduction

- Bilingue et multilingue
 - Les usagers peuvent exprimer leurs demandes dans une langue et comprendre une autre;
 - Ils peuvent écrire leur besoin d'information dans leur propre langue et comprendre une autre (e.g., texte très court, QA, statistiques et autre information factuelle, image, graphique, musique);
 - Ils désirent avoir une idée approximative du contenu (et, au besoin, obtenir une traduction manuelle des documents souhaités plus tard);
 - plus important sur le Web (voyage, réservation d'hôtel) (mais les consommateurs préfèrent avoir les informations précises dans leur langue, particulièrement pour des biens/services moins standardisés).

EARIA'06 - 7 -

Campagne d'évaluation

- TREC (trec.nist.gov)
 - TRECs 3-5: Espagnol
 - TRECs 5-6: Chinois (simplifié, GB)
 - TRECs 6-8: Multilingue (EN, DE, FR, IT)
 - TREC-9: Chinois (traditionnel, BIG5)
 - TRECs 10-11: Arabe

[Harman 2005]

EARIA'06 - 8 -

Campagne d'évaluation

- CLEF (www.clef-campaign.org)
 - Débute en 2000 avec EN, DE, FR, IT
 - 2001-02: EN, DE, FR, IT, SP, NL, FI, SW
 - 2003: DE, FR, IT, SP, SW, FI, RU, NL
 - 2004: EN, FR, RU, PT
 - 2005-06: FR, PT, HU, BG
 - 2007: HU, BG, CZ, RO(?)
 - Evaluation monolingue, bilingue et multilingue
 - Autres pistes: domaine spécifique, interactive, document audio (2002 →), Image-CLEF (2003 →), QA(2003 →), Web(2005 →), GeoCLEF (2005 →) [Braschler & Peters 2004]

EARIA'06 - 9 -

Campagne d'évaluation

Les requêtes sont disponibles en plusieurs langues
(CLEF 2005)

- EN: Nestlé Brands
- FR: Les Produits Nestlé
- PT: Marcas da Nestlé
- HU: Nestlé márkák
- BG: Продуктите на Нестле
- EN: Italian paintings
- FR: Les Peintures Italiennes
- PT: Pinturas italianas
- HU: Olasz (itáliai) festmények
- BG: Италиански картини

EARIA'06 - 10 -

Campagne d'évaluation

- NTCIR (research.nii.ac.jp/ntcir/)
 - Début en 1999: EN, JA
 - NTCIR-2 (2001): EN, JA, ZH (traditionnel)
 - NTCIR-3 (2002), NTCIR-4 (2004), et NTCIR-5 (2005): EN, JA, KR, ZH (traditionnel), brevets (JA), QA (JA), Web (.jp), résumé automatique
 - NTCIR-6 (2007): JA, KR, ZH (traditionnel), évolution des opinions, brevets, résumé automatique

EARIA'06 - 11 -

Plan

- Introduction
- **RI monolingue**
- Problèmes de traduction
- Stratégies de traduction (RI bilingue)
- RI multilingue

EARIA'06 - 12 -

Le problème ...

<TOPIC>

<TITLE>時代華納，美國線上，合併案，後續影響</TITLE>

<DESC> 查詢時代華納與美國線上合併案的後續影響。</DESC>

<NARR>

<BACK>時代華納與美國線上於2000年1月10日宣佈合併，總市值估計為3500億美元，為當時美國最大宗合併案。</BACK>

<REL>評論時代華納與美國線上的合併對於網路與娛樂媒體事業產生的影響為相關。敘述時代華納與美國線上合併案的發展過程為部分相關。內容僅提及合併的金額與股權結構轉換則為不相關。</REL>

</NARR>

<CONC>時代華納，美國線上，李文，Gerald Levin，合併案，合併及採購，媒體業，娛樂事業</CONC>

</TOPIC>

EARIA'06 - 13 -

RI monolingue

- Quelques exemples
 - Strč prst skrz krk
 - Mitä sinä teet?
 - Mam swoją książkę
 - Nem fáj a fogad?
 - Er du ikke en rigtig nordmann?
 - Добре дошли в България!
 - Fortuna caeca est
 - نهار سعيد

EARIA'06 - 14 -

RI monolingue

- Alphabets
 - Latin (26)
 - Cyrillique (33)
 - Arabe (28), Hébreux (22 consonnes)
 - Autres langues d'Asie: Hindi, Thai
- Syllabaires
 - Japon: Hiragana (46) における
Katakana (46) フランス
 - Corée: Hangul (8 200) 정보검색시스템
- Idéogrammes
 - Chine (13,000/7,700) 中国人, Japon (8,800) ボ紛争

EARIA'06 - 15 -

RI monolingue

- La représentation dans d'autres langues
 - ASCII est limité à 7 bits
 - Windows, Macintosh, BIG5, GB, EUC-JP, EUC-KR, ...
 - ISO-Latin-1 (ISO 8859-1 West European), Latin-2 (East European), Latin-3 (South European), Latin-4 (North European), cyrillique (ISO-8859-5), arabe (ISO-8859-6), grec (ISO-8859-7), hébreux (ISO-8859-8), ...
 - Unicode (UTF-8, voir www.unicode.org)

EARIA'06 - 16 -

RI monolingue

- Les entrées / sorties
 - Comment introduire un texte / imprimer un document (ou phrase) avec des idéogrammes ?
 - Yudit (www.yudit.org)
 - écriture de droite à gauche (Arabic)
 - les caractères cyrilliques
- Outils
 - Quel sera le résultat d'un `wc`, `grep`?
 - Quel sera le résultat d'un `sort` avec des mots en japonais ?

EARIA'06 - 17 -

RI monolingue (segmentation)

- Qu'est-ce qu'un mot / unité d'indexation ?
 - Les mots composés (*worldwide*, *handgun*) sont utilisés très fréquemment dans certaines langues (DE, NL, FI, HU, BG)
Le français soulève aussi ses propres problèmes (“chemin de fer”, “pomme de terre”)
 - En DE: “Bundesbankpräsident” =
“Bund” + es + “Bank” + “Präsident”
fédéral banque PDG
mais le défi est le fait que plusieurs formes peuvent coexister.
Ainsi “ComputerSicherheit” peut aussi apparaître comme “die Sicherheit mit Computern”
 - Décomposition automatique en mots (+23% en MAP, requêtes T, +11% requêtes TD, [Braschler & Ripplinger 2004])
 - Même procédure pour d'autres langues (FI, HU, NL).

EARIA'06 - 18 -

Décomposition

- Comment retrouver les unités lexicales dans un mot composé (par exemple en DE “Atomtests,” “Wintersport”)?
- S'appuyer sur les trigrammes très peu fréquent ou impossible (e.g., la séquence “fff” est impossible, donc “Schiffahrt” est composé comme “Schiff+fahrt”) (UniNE à CLEF-2002)
- Utiliser une liste de mots (avec leur fréquence dans un corpus) (Berkley à CLEF-2002)

computer	2 452	port	1 091
computers	79	ports	2
sicherheit	6 583	sport	1 483
sicher	4 522	winter	1 643
bank	9 657	winters	148
bund	7 032	wintersport	44
bundes	2 884	wintersports	2
bundesbank	1 453		
präsident	24 041		

EARIA'06 - 19 -

RI monolingue (segmentation)

- Mais cela devient “indispensable” en ZH

我不是中国人
我 不 是 中 国 人
Je pas être Chinois

- Diverses stratégies de décomposition (automatique) ont été proposées
(l'appariement le plus long, information mutuelle, programmation dynamique, analyseur morphologique, voir MandarinTools (www.mandarintools.com))

EARIA'06 - 20 -

RI monolingue (segmentation)

Cela devient un peu plus simple en JA

コソボ紛争におけるNATOの攻撃と

Kanji (idéogrammes chinois)	42.3 %
Hiragana (e.g., dans, de, ...)	32.1 %
Katakana (e.g., フランス)	7.9 %
Romaji (notre alphabet)	7.6 %
...autres	10.1 %

voir l'analyseur morphologique Chasen (chasen.aist-nara.ac.jp)

statistiques sur le corpus NTCIR-5

EARIA'06 - 21 -

RI monolingue (segmentation)

Pour le coréen, la même expression peut s'écrire à l'aide de plusieurs constructions différentes

정보 (information) 검색 (recherche) 시스템 (système)
정보검색 (information recherche) 시스템 (système)
정보 (information) 검색시스템 (recherche système)
정보검색시스템

voir l'analyseur morphologique Hangul (nlp.kookmin.ac.kr)

EARIA'06 - 22 -

RI monolingue

- Approche indépendante de la langue
l'indexation par *n*-gramme [McNamee & Mayfield 2004]
 - segmentation automatique de chaque phrase
 - deux formes possibles
“The White House”
 - “The “, “he W”, “h Wh”, “ Whi”, “Whit”, “hite”, ...
 - ou
 - “the“, “whit”, “hite”, “hous”, “ouse”
 - Présente une approche efficace si l'on connaît pas (ou peu) la langue
 - une stratégie d'indexation classique pour JA, ZH ou KR

EARIA'06 - 23 -

RI monolingue

Une phrase en chinois

我不是中国人

unigrammes

我 不 是 中 国 人

bigrammes

我 不 不 是 是 中 中 国 国 人 人

unigrammes et bigrammes

我, 不, 是, 中, 国, 人, 我, 不, 不是, 是中, 中国, 国人

mots (MTSeg)

我 不 是 中国人

EARIA'06 - 24 -

RI monolingue

ZH: unigramme & bigramme > mot (MTool) \approx bigramme

L'indexation par n -gramme meilleure qu'une segmentation (automatique) (MTool) [Abdou & Savoy 2006]

valeur de référence en gras, différence stat. significative soulignée

JA: unigramme & bigramme \approx mot (Chasen) = bigramme

Chinois (T) NTCIR-5	Unigram.	Bigram. (réf.)	mot (MTool)	Uni + bigram
PB2	0,2774	0,3042	0,3246	<u>0,3433</u>
LM	0,2995	0,2594	0,2800	0,2943
Okapi	0,2879	0,2995	0,3231	<u>0,3321</u>
<i>tf idf</i>	<u>0,1162</u>	0,2130	<u>0,1645</u>	0,2201

EARIA'06 - 25 -

RI monolingue

KR: bigramme \approx HAM > unigramme [Abdou & Savoy 2006]

L'indexation par n -gramme toujours meilleure (pas statistiquement [Abdou & Savoy 2006])

Coréen (T) NTCIR-5	Unigram.	Bigram.	Décompos. (HAM)
PB2	<u>0,2378</u>	0,3729	0,3659
LM	<u>0,2120</u>	0,3310	0,3135
Okapi	<u>0,2245</u>	0,3630	0,3549
<i>tf idf</i>	<u>0,1568</u>	0,2506	0,2324

RI monolingue

- Accents
 - différent d'une langue à l'autre (“résumé”, “Äpfel”, “leão”);
 - permettent parfois de spécifier le sens (e.g., “tache” ou “tâche”);
 - mais souvent le sens entre deux formes différentes est relié (e.g., “cure” et “curé”, “Apfel” et “Äpfel”);
 - très souvent les accents sont supprimés par le système de dépistage (les différences en MAP ne sont pas statistiquement significatives).

EARIA'06 - 27 -

RI monolingue

- Normalisation / noms propres
 - homophones des noms propres. Par exemple, Stephenson (inventeur de la machine à vapeur), et Stevenson (l'auteur) possèdent la même prononciation dans les langues japonaise, chinoise ou coréenne. Les deux noms sont donc orthographiés de la même manière.
 - Le nom peut varier dans une même langue ou entre les langues (Gorbachev, Gorbacheff, Gorbachov)
 - Pas de règle très stricte (par exemple, en FR “cowboy” et “cowboy” “véto” et “veto” ou “eczéma” et “exéma” (comme en anglais avec “color” et “colour”, etc.). En DE: plusieurs réformes de l'orthographe.

EARIA'06 - 28 -

RI monolingue

- Les listes de mots outils (antédictionnaires)
 - Mot fréquent et peu porteur de sens (+ pronoms, prépositions, conjonctions)
 - Mais la séparation n'est pas toujours claire ("or" en FR)
et avec les accents ("été" mais "ete" n'existe pas).
 - La liste complète peut dépendre du système (e.g., un système de question/réponse requiert les pronoms interrogatifs)
 - Peut également être dépendant de la requête (on élimine les mots apparaissant souvent dans des requêtes différentes)
(voir ThomsonRL à NTCIR-4)

EARIA'06 - 29 -

RI monolingue (raciniseur)

- raciniseur (mots & règles)
 - Inflexions
 - nombre (sing / plural), chat, chats
 - genre (femi / masc), chatte,
 - personne, temps, voix, aimerais, aimées
 - reste simple en EN ('-s', '-ing', '-ed')
 - Dérivations
 - former des nouveaux mots (et changer la catégorie grammaticale POS)
 - en EN '-ably', '-ment', '-ship'
 - admit → {admission, admittance, admittedly}
 - en FR '-able', '-ment', '-teur'

EARIA'06 - 30 -

RI monolingue (raciniseur)

Raciniseur léger en FR [Savoy 2004]

pour par exemple “barons” → “baron”,
“baronnes” → “baron”)

Pour mot de six lettres ou plus, faire

si la finale est ‘-aux’ alors remplacez ‘-aux’ par ‘-al’,

si la finale est ‘-x’ alors éliminez ‘-x’,

si la finale est ‘-s’ alors éliminez ‘-s’,

si la finale est ‘-r’ alors éliminez ‘-r’,

si la finale est ‘-e’ alors éliminez ‘-e’,

si la finale est ‘-é’ alors éliminez ‘-é’,

si les deux dernières lettres sont identiques, alors
éliminez la dernière

EARIA'06 - 31 -

RI monolingue (raciniseur)

Pour les langues germaniques, un raciniseur léger est un peu plus complexe

- Plusieurs suffixes peuvent désigner le pluriel (voire le recours à des accents)
“Motor”, “Motoren”; “Jahr”, “Jahre”;
“Apfel”, “Äpfel”; “Haus”, “Häuser”
- Les différents cas grammaticaux peuvent impliquer des suffixes
(e.g., génitif avec ‘-es’ “Staates”, “Mannes”)
mais aussi après les adjectifs
 (“einen guten Mann”)
- Les mots composés
(“Lebensversicherungsgesellschaftsangestellter”
= vie + assurance + société + employé)

EARIA'06 - 32 -

RI monolingue (raciniseur)

La famille finno-ouralienne possède de nombreux cas (18 en HU)

ház	nominatif (maison)
házat	accusatif singulier
házakat	accusatif pluriel
ház ^{al}	“avec” (instrumental)
ház ^{on}	“sur” (superessif)
házamat	ma + accusatif singulier
házamiat	ma + accusatif + pluriel

- en FI, la racine même se modifie (e.g., “matto”, “maton”, “mattoja” (tapis))
Il semble que pour le FI on doive recourir à une analyse morphologique plus poussée (voir Hummingbird, CLEF 2004, p. 221-232)
- + des mots composés (“internetfüggök”, “rakkauskirje”)

EARIA'06 - 33 -

RI monolingue (raciniseur)

Différentes stratégies pour construire un raciniseur

- Appliquer les règles de la grammaire (les plus fréquentes pour le moins) (on peut ignorer “fou” → “folle”)
- Recourir à un dictionnaire pour réduire le taux d’erreur [Krovetz 1993], [Savoy 1993]
- Consulter les usages faits dans la langue (contrairement aux règles de grammaire) [Xu & Croft 1998]
- “Ignorer” le problème et procéder par une indexation par *n*-gramme
e.g., “bookshop” ? “book”, “ooks”, “oksh”
et cela s’avère efficace pour le ZH, JA, KR ...
[McNamee & Mayfield 2004]

EARIA'06 - 34 -

RI monolingue (raciniseur)

- Evaluations
- Les expériences les plus nombreuses sont dans les actes de CLEF / TREC
- Autres évaluations dans [Savoy 2006]
- Les tendances générales (MAP)
 - Avec enracineur (stemmer) > rien faire
 - Différences entre enracineurs ne sont pas souvent statistiquement significatives
 - Enracineur léger noms + adjectifs peut être meilleur ou apporter une performance similaire à un enracineur plus agressif (erreur avec les formes verbales très nombreuses?)
 - Pour les langues d'Extrême Orient: pas clair
JA: éliminer les caractères Hiragana

EARIA'06 - 35 -

RI monolingue (raciniseur)

Le raciniseur peut connaître des problèmes et générer des d'erreurs

Dans la requête (HU)

"internetfüggok" (dépendance à internet – personne
«függ» est le verbe – la racine attendue -)

Dans les documents pertinents

"internetfüggoség" (dépendance) → "internetfüggoség"

"internetfüggoséggel" ("avec") → "internetfüggoség"

"internetfüggoségben" ("dans") → "internetfüggoség"

→ le raciniseur échoue

EARIA'06 - 36 -

RI monolingue (raciniseur)

Corpus CLEF-2005, requêtes T

FR (T)	none	UniNE	léger '-s'	Porter
Okapi	0,2260	<u>0,3045</u>	<u>0,2858</u>	<u>0,2978</u>
GL2	0,2125	<u>0,2918</u>	<u>0,2739</u>	<u>0,2878</u>
Lnu-ltc	0,2112	<u>0,2933</u>	<u>0,2717</u>	<u>0,2808</u>
dtu-dtn	0,2062	<u>0,2780</u>	<u>0,2611</u>	<u>0,2758</u>
<i>tf-idf</i>	0,1462	<u>0,1918</u>	<u>0,1807</u>	<u>0,1758</u>

EARIA'06 - 37 -

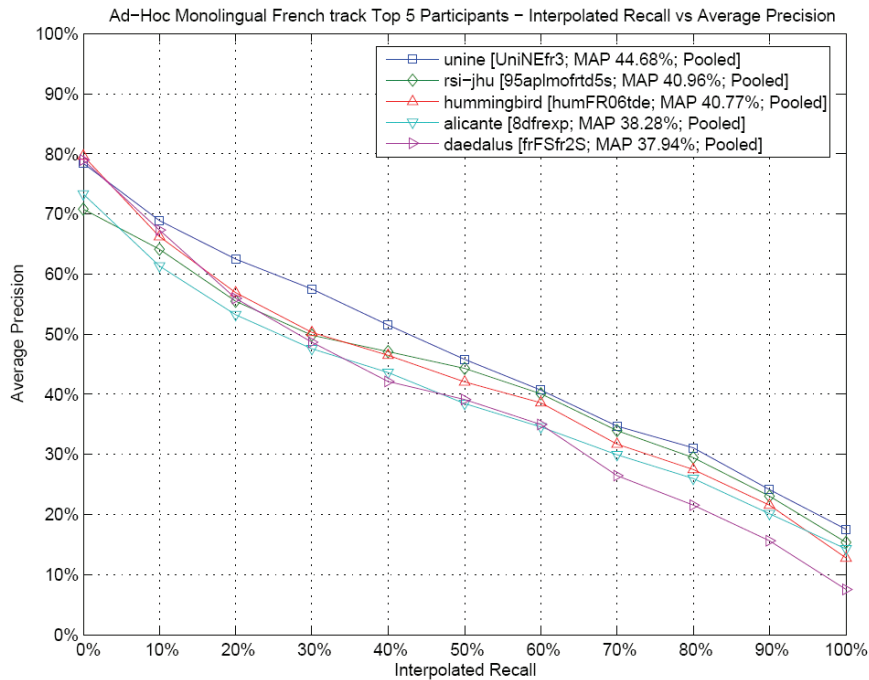
RI monolingue (raciniseur)

Corpus CLEF-2005, requêtes T

FR (T)	none	UniNE	léger '-s'	Porter
Okapi	<u>0,2260</u>	<u>0,3045</u>	0,2858	0,2978
GL2	<u>0,2125</u>	<u>0,2918</u>	0,2739	0,2878
Lnu-ltc	<u>0,2112</u>	<u>0,2933</u>	0,2717	0,2808
dtu-dtn	<u>0,2062</u>	0,2780	0,2611	0,2758
<i>tf-idf</i>	<u>0,1462</u>	0,1918	0,1807	0,1758

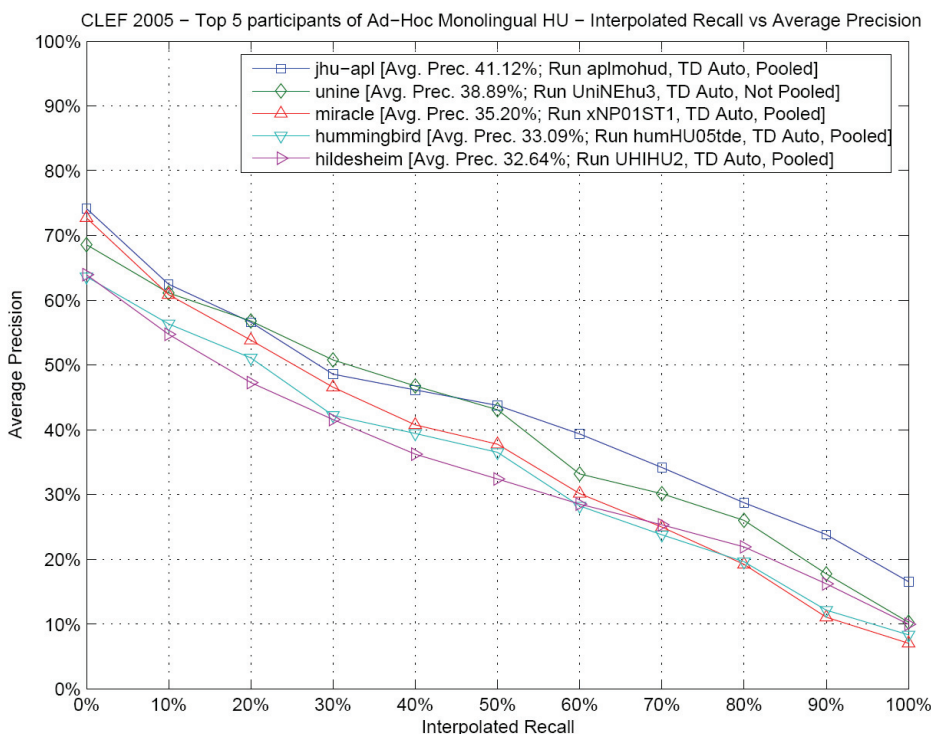
EARIA'06 - 38 -

RI monolingue (CLEF 2006)



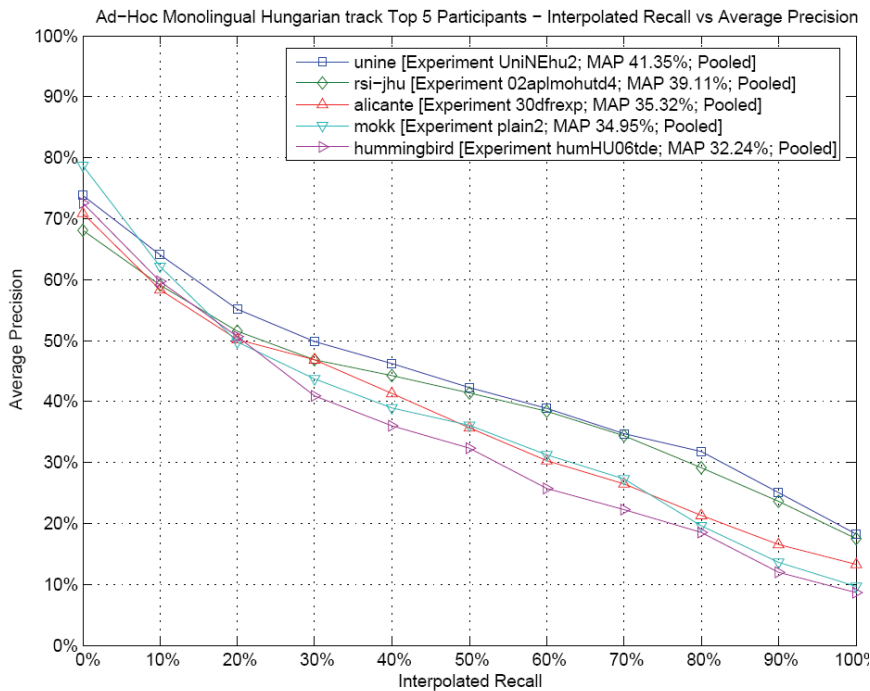
- FR, une langue connue
- Différences de performance (MAP) sont relativement faibles entre les cinq meilleures
- Plusieurs stratégies différentes tendent à produire des MAP similaires

RI monolingue (CLEF 2005)



- HU, nouvelle langue
- n -gramme propose la meilleure performance
- Mais on espère une amélioration via des approches qui tiennent compte de la langue.

RI monolingue (CLEF 2006)



Mais cela peut
changer avec le
temps

EARIA'06 - 41 -

Plan

- Introduction
- RI monolingue
- **Problèmes de traduction**
- Stratégies de traduction (RI bilingue)
- RI multilingue

EARIA'06 - 42 -

Problème de traduction

- “non verbum e verbo, sed sensum exprimere de sensu”
- “horse” = “cheval”?
 - oui (si on pense à une animal à quatre pattes)
“horse-race” = course de chevaux
 - Oui dans le sens, pas vraiment dans la forme
“horse-show” = “concours hippique”
“horse-drawn” = “hippomobile”
 - Sens différent et traduction différente
“horse-fly” = “taon”
“horse sense” = “gros bon sens”
“to eat like a horse” = “manger comme un loup”

EARIA'06 - 43 -

Problème de traduction

- Emprunt
“full-time” → “temps plein”(*)
- Calque
“igloo” → “iglou”
- Traduction mot à mot
 - “a lame duck Congressman” ? “canard boiteux”(*)
 - Faux amis
“Requests of Quebec” = “Demandes du Québec”
“Demands of Quebec” = “Exigences posées par le Québec”
- Translation = équivalence dans le sens (et forcément dans la forme “Yield” = “Priorité à gauche” → “Cédez”)

EARIA'06 - 44 -

Traduction automatique

- “Tainted-Blood Trial”
 - Manually “L'affaire du sang contaminé”
 - Systran “Épreuve De Corrompu - Sang”
 - Babylon “entacher sang procès”
- “Death of Kim Il Sung”
 - Manually “Mort de Kim Il Sung”
 - Systran “La mort de Kim Il chantée”
 - Babylon “mort de Kim Il chanter”
 - Babylon “Tod von Kim llinium singen ”
- “Who won the Tour de France in 1995?”
 - Manually “Qui a gagné le tour de France en 1995”
 - Systran “Organisation Mondiale de la Santé, le, France 1995 ”

EARIA'06 - 45 -

Plan

- Introduction
- RI monolingue
- Problèmes de traduction
- **Stratégies de traduction (RI bilingue)**
- RI multilingue

EARIA'06 - 46 -

Traduction automatique

- La traduction automatique ajoute de l'ambiguïté
 - On dispose de plusieurs traductions pour chaque mot
 - Utiliser une probabilité de traduction correcte (comment ?)
 - L'expansion de la requête peut aider
- On a besoin de ressources supplémentaires pour chaque langue
 - Dictionnaire bilingue ou multilingue (ou liste de mots)
 - Liste de noms propres
 - Corpus parallèle
 - Corpus comparable (thème, temps, culture)
 - Système de traduction automatique (MT)
- Les approches statistiques dominant le débat [Gao & Nie, 2006]

EARIA'06 - 47 -

Traduction automatique

- On ignore la traduction !

On considère qu'une phrase dans une langue est simplement mal orthographiée dans l'autre langue. On a simplement besoin d'un correcteur orthographique (e.g., Cornell à TREC-6, Berkeley à NTCIR-5)
- Traduction uniquement des requêtes
 - moins / peu coûteux
- Traduction uniquement des documents
 - à effectuer avant la recherche (offline)
- Traduction des requêtes et des documents
 - coûteux mais efficace
- La qualité de réponse varie entre 50 à 75 % de celle obtenue en interrogation unilingue (TREC-6) voire de 80 % à 100 % (CLEF 2005-2006)

EARIA'06 - 48 -

Traduction automatique

- Dictionnaire bilingue (multilingue) en-ligne (MRD)
 - redonne habituellement plus qu'une traduction possible (les prendre toutes ? La première seulement ? Avec le même poids ?)
 - problème des mots qui ne sont pas dans le dictionnaire (e.g., les noms propres) [Zhang *et al.* 2005]
 - peut être limité à une simple liste de mots
- Traduction automatique par machine (MT)
 - plusieurs systèmes disponibles (gratuitement)
 - la qualité (et l'interface) varie d'un système à l'autre
- Modèles statistiques [Nie *et al.* 1999], [Gao & Nie 2006]
 - plusieurs modèles statistiques ont été proposés
 - voir le projet MBOI au rali.iro.umontreal.ca/

EARIA'06 - 49 -

Traduction automatique

- Une expansion de la requête avant la traduction peut parfois aider
 - mais cela risque aussi d'être une difficulté pour un système de traduction automatique
- Une expansion de la requête après la traduction
 - cela améliore la MAP
- Corpus parallèle (ou comparable)
 - peuvent être difficile à obtenir
 - les différences thématiques, de temps ou de culture jouent un rôle
 - on peut recourir au Web ou à des ressources de meilleure qualité (e.g. Wikipedia, journaux)
- La structuration d'une requête peut aider le système de dépistage [Hedlund *et al.* 2004]
- Une meilleure traduction des syntagmes nominaux est importante
- Les campagnes d'évaluation utilisent souvent des noms propres dans les requêtes
 - Il peut s'avérer utile de les traiter

EARIA'06 - 50 -

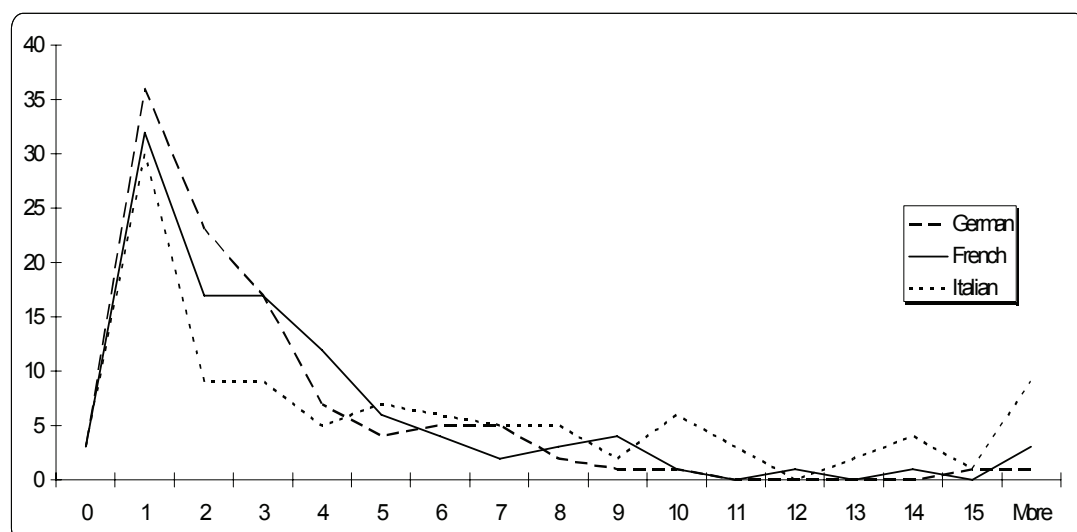
Différence culturelle

- Le même concept peut posséder différentes traductions selon la régions ou le pays
 - E .g. “Mobile phone”
 - « *Natel* » en Suisse
 - « *Cellulaire* » au Québec
 - « *Téléphone portable* » en France
 - « *Téléphone mobile* » en Belgique

EARIA'06 - 51 -

Traduction

Le nombre de traductions proposées par un dictionnaire bilingue (Babylon) reste habituellement faible



EARIA'06 - 52 -

Stratégies de traduction

Exemple de titre

- “Final Four Results”
 - en FR: “final quatre résultat” (Babylon)
au lieu de “Résultats des demi-finales”
 - en DE: “Resultate Der Endrunde Vier “ (Systran)
au lieu de “Ergebnisse im Halbfinale”
- “Renewable Power ”
 - en FR, au lieu de “Energie renouvelable”
“Puissance Renouvelable”
“renouvelable pouvoir”
- “Mad Cow Dease ”
 - en FR, au lieu de “maladie de la vache folle”
“fou vache malade”
et l’*enracineur* n’a pas réduit sous la même racine les termes “fou” et
“folle”

EARIA'06 - 53 -

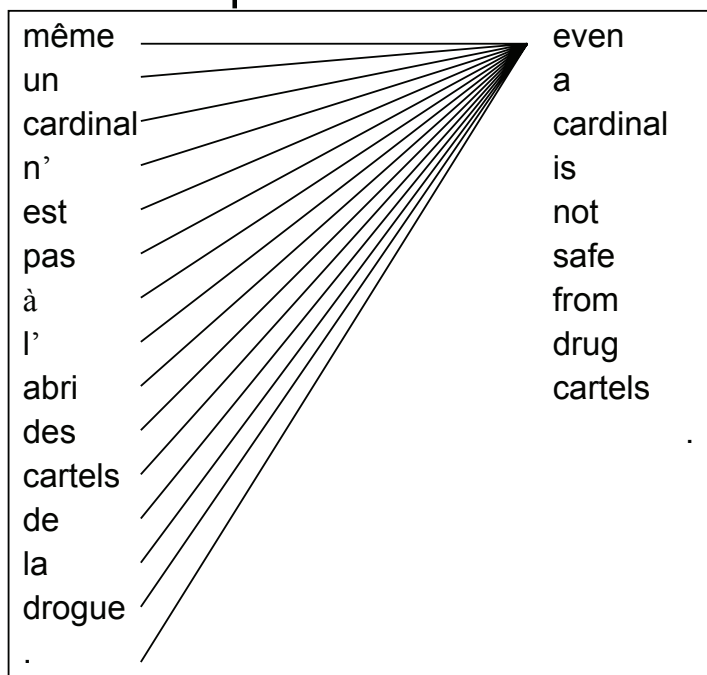
Stratégies de traduction

- Approche statistique (de l’EN au FR)
- $P[e_j|f_i]$ est estimé via un corpus d’entraînement parallèle, après alignement (automatique) des paragraphes puis des phrases [Gale & Church, 1993]
- Pas d’information syntaxique ou de position (model 1 d’IBM, [Brown *et al.*, 1993])
- Procédure:
 - Input = deux ensembles de texte parallèle
 - Alignement des phrases A : $E_k \leftrightarrow F_l$
 - Affectation des probabilités initiales: $t[e_j|f_i, A]$
 - Expectation Maximization (EM): $P[e_j|f_i, A]$
 - Résultat final: $P[e_j|f_i] = P[e_j|f_i, A]$

EARIA'06 - 54 -

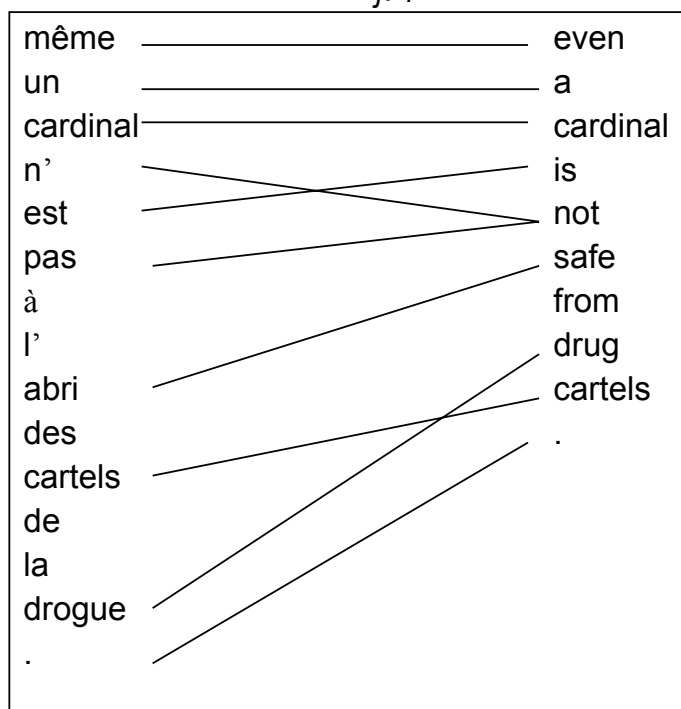
Stratégies de traduction

Affectation des probabilités initiales $t[e_j|f_i, A]$



Stratégies de traduction

Application du principe EM: $P[e_j|f_i, A]$



Stratégies de traduction

Avec un corpus parallèle [Gale & Church 1991]

- Exemple extrait du système MBOI (rali.iro.umontreal.ca/)
- Requête “database system”
 - en FR: “(données^0.29472154 base^0.20642714 banque^0.037418656)”
 - “système de bases de données”

EARIA'06 - 57 -

Traduction

Une meilleure traduction ne redonne pas forcément une meilleure performance en recherche !

Outil de traduction	Requête	MAP
EN (original)	U.N./US Invasion of Haiti. Find documents on the invasion of Haiti by U.N./US soldiers.	
Reverso	Invasion der Vereinter Nationen Vereinigter Staaten Haitis. Finden Sie Dokumente auf der Invasion Haitis durch Vereinte Nationen Vereinigte Staaten Soldaten.	40,07
Free	U N UNS Invasion von Haiti. Fund dokumentiert auf der Invasion von Haiti durch U N UNS Soldaten	72,14

EARIA'06 - 58 -

Traduction

Comparaison de 11 traductions manuelles du lot de requête en EN (T) [Savoy 2003]

- grande variabilité entre traducteurs
- les traductions fournies par CLEF sont bonnes
(différences statistiquement significatives sont soulignées, test bilatéral, $\alpha=5\%$)

	CLEF	moyenne	max	min
Okapi	0,4162	<u>0,3516</u>	0,4235	0,2929
<i>tf idf</i>	0,2502	<u>0,1893</u>	0,2416	0,0261
binary	0,2285	<u>0,1662</u>	0,2151	0,0288

EARIA'06 - 59 -

Traduction

Les requêtes originales sont écrites en EN (T, Okapi, CLEF-2000)

- Traduction automatique par Systran
- par Babylon (en prenant uniquement la 1ère traduction)
- par concaténation des deux traductions

	Manual	Systran	Babylon	Combined
FR mot	0,4162	<u>0,2964</u> (-29 %)	<u>0,2945</u> (-29 %)	<u>0,3314</u> (-20 %)
DE 5-gram	0,3164	<u>0,2259</u> (-29 %)	<u>0,1739</u> (-45 %)	0,2543 (-20 %)
IT mot	0,3398	<u>0,2079</u> (-39 %)	<u>0,1993</u> (-41 %)	<u>0,2578</u> (-24 %)

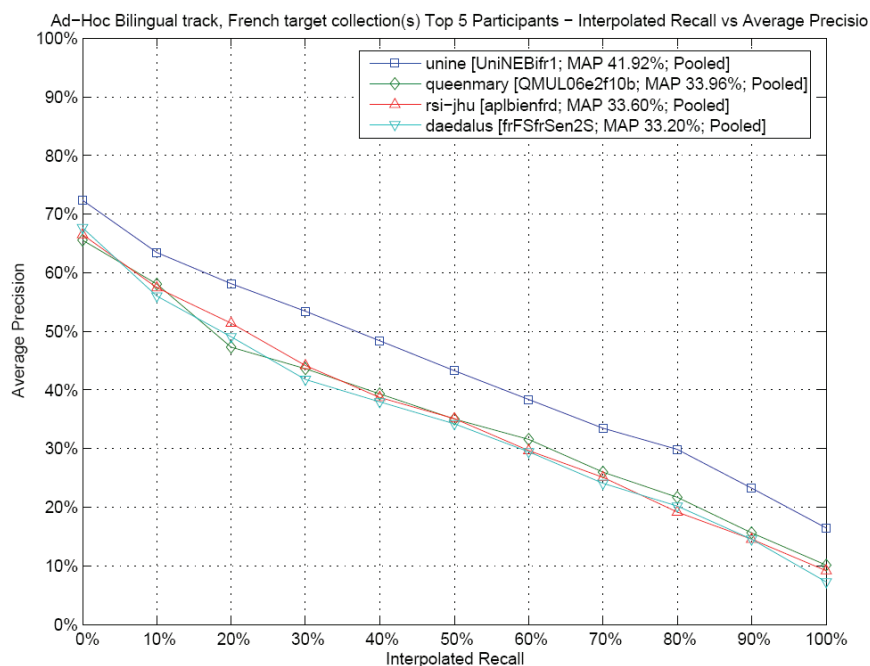
EARIA'06 - 60 -

Evaluation

- Différentes situations sont possibles
 - Les langues peuvent disposer d'un nombre variable d'outils de traduction / de corpus parallèle et comparable / d'outils morphologique / expériences en RI
 - Des langues naturelles peuvent être plus faciles que d'autres (d'un point de vue RI).
- La comparaison entre les expériences bilingues et monolingues n'est pas toujours possible
 - Certaines équipes participent seulement à une piste
 - Pour les deux expériences, le moteur de recherche n'est pas forcément le même
 - Les paramètres ne sont pas les mêmes entre les expériences faites en monolingue et celles faites en bilingue

EARIA'06 - 61 -

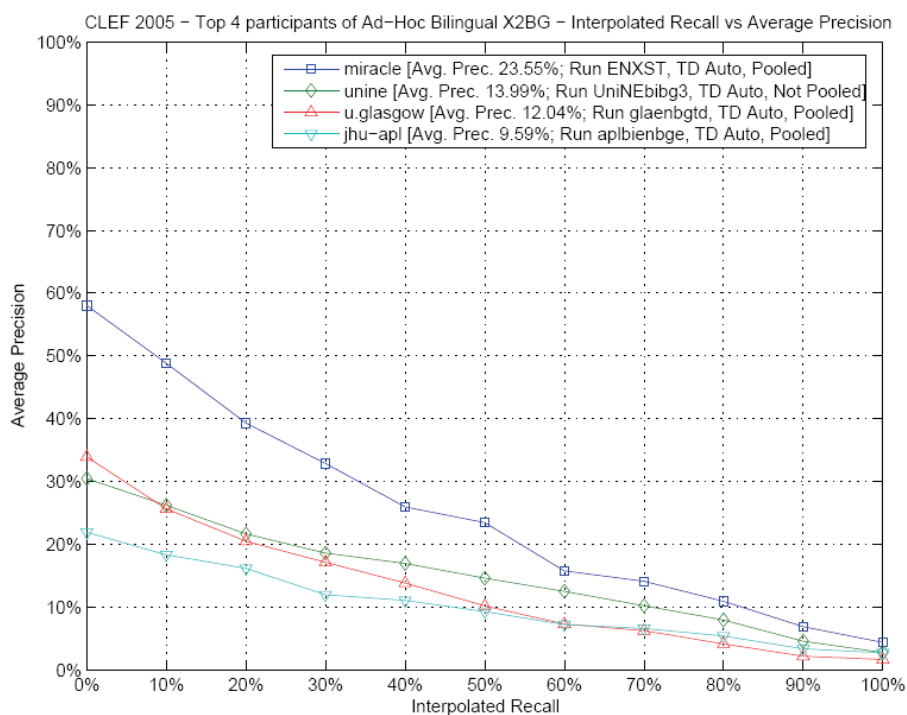
CLIR (CLEF-2006 X ? FR)



- FR est une langue plus connue (en RI)
- Plusieurs outils de traduction existent
- Piste proposée pendant plusieurs années
- Meilleur MAP mono: 0,4468 (diff=-6,2%)
- Peu de différence entre le 2^e et le 4^e.

EARIA'06 - 62 -

CLIR (CLEF-2005 X ? BG)



BG est une nouvel langue
Peu d'outils de traduction disponibles
Première année
Meilleur MAP mono:
0,3203 (diff=-26,5%)
La qualité de l'outil de traduction explique la différence entre les deux premiers

EARIA'06 - 63 -

Plan

- Introduction
- RI monolingue
- Problèmes de traduction
- Stratégies de traduction (RI bilingue)
- **RI multilingue**

EARIA'06 - 64 -

Stratégie multilingue en IR

- Créé un index multilingue
(voir Berkeley TREC-7)
 - on construit un index pour tous les docs (quelque soit la langue)
 - on traduit la requête soumise dans toutes les langues
 - recherche dans un index multilingue et on retourne directement une liste de documents écrits dans plusieurs langues
- Créé un index commun sur la base des documents traduits automatiquement (DT) (voir Berkeley CLEF-2003)
 - Sélection d'un langue pivot (EN!) (Berkeley à CLEF-2003)
 - Traduction des documents et requêtes dans cette langue
 - Recherche dans un index (volumineux mais écrit dans une seule langue) et on retourne une liste unique de documents

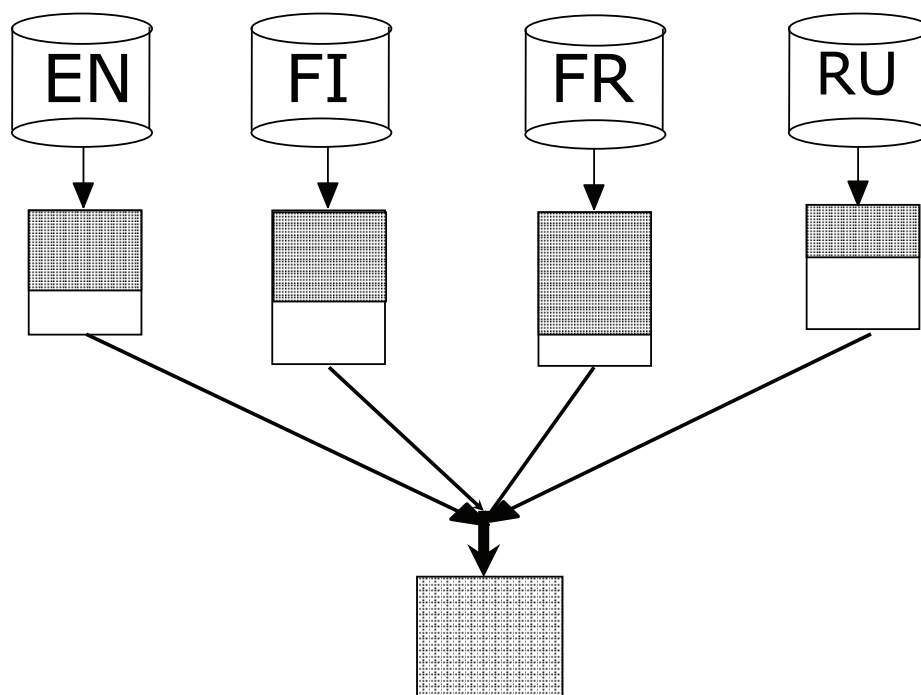
EARIA'06 - 65 -

Stratégie multilingue en IR

- La traduction de la requête (QT)
 - Traduire automatiquement la requête dans plusieurs langues
 - Effectuer une recherche séparée dans chaque langue
 - Fusionner les résultats obtenus dans les diverses langues
- Mélange QT et DT (Berkeley à CLEF 2003, Eurospider à CLEF 2003) [Braschler 2004]
- Pas de traduction
 - Possible dans des paires de langues voisines ou ayant la même écriture (JA-ZH)
 - Application multilingue limitée (e.g., noms propres, lieux et noms géographiques)

EARIA'06 - 66 -

Stratégie multilingue en IR (QT)



EARIA'06 - 67 -

RI multilingue

Le problème de la fusion

1	EN120	1.2
2	EN200	1.0
3	EN050	0.7
4	EN705	0.6
...		

1.	FR043	0.8
2.	FR120	0.75
3.	FR055	0.65
4.	...	

1	RU050	6.6
2	RU005	6.1
3	RU120	3.9
4	...	

EARIA'06 - 68 -

RI multilingue

- Voir également la “recherche distribuée”
- A chacun son tour (*round-robin*)
- Fusion par le score brut (*raw-score merging*)
 $Score_j(D_i)$ score du document obtenu par le système j
 $RSV(D_i)$ score final du document

$$RSV(D_i) = \sum_{j=1}^k Score_j(D_i)$$

- Normalisation (e.g, par le score du premier doc. = max)

$$RSV(D_i) = \sum_{j=1}^k Score'_j(D_i)$$

$$\text{avec } Score'_j(D_i) = \frac{Score_j(D_i)}{ScoreMax_j}$$

RI multilingue

- A chacun son tour biaisé
on sélectionne plus qu'un document dans les listes en fonction de la qualité de ces listes)

- Z-score

On calcule la moyenne et l'écart-type

$$RSV(D_i) = \sum_{j=1}^k Score'_j(D_i)$$

$$\text{avec } Score'_j(D_i) = \frac{(Score_j(D_i) - \mu_j) + \delta_j}{\sigma_j}$$

- Régression logistique [Le Calvé 2000], [Savoy 2004]

$$Score'_j(D_i) = \frac{1}{1 + e^{-[\alpha_j + \beta_{1j} \cdot \ln(rank(D_i)) + \beta_{2j} \cdot RSV(D_i)]}}$$

RI multilingue

Cond. A le meilleur système par langue (CLEF 2004) (dont on fusionne des systèmes différents)

Cond C le même système pour toutes les langues

EN → {EN, FR, FI, RU}	Condition A	Condition C
A chacun son tour	0,2386	0,2358
Fusion par le score brut	0,0642	0,3067
Normalisation (max)	0,2899	0,2646
A chacun son tour biaisé	0,2639	0,2613
Z-score	0,2669	0,2867
Régression logistique	0,3090	0,3393

EARIA'06 - 71 -

RI multilingue

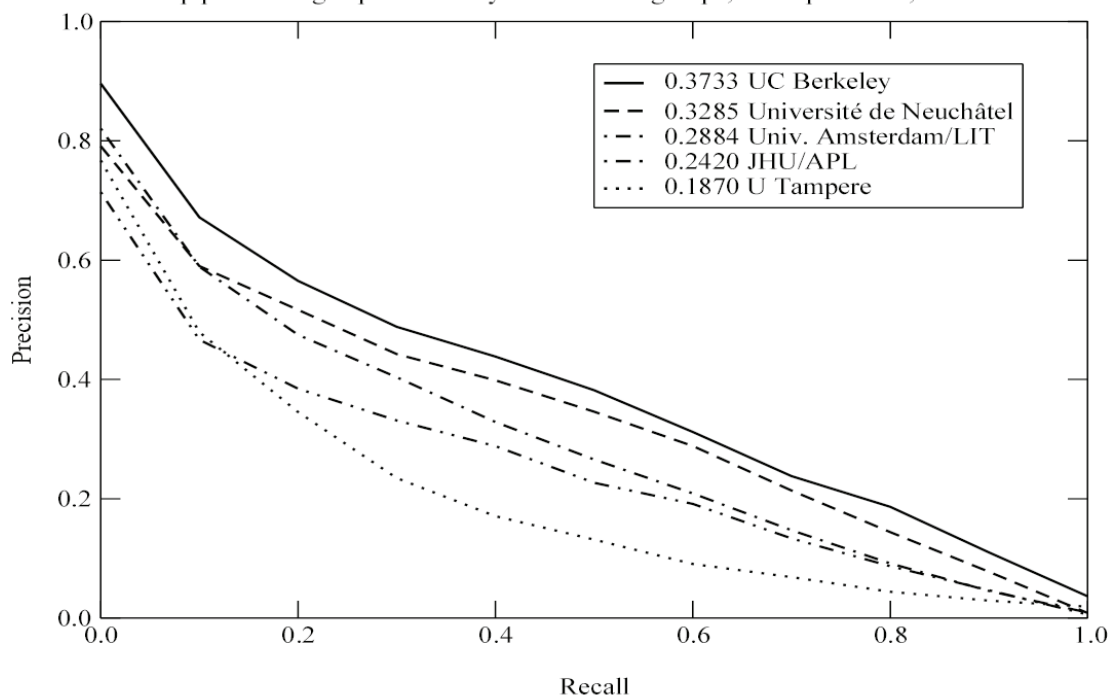
- Traduction des requêtes et fusion
 - La régression logistique fonctionne bien (apprentissage sur les données CLEF 2003, évaluation sur CLEF 2004)
 - La normalisation est aussi une bonne stratégie (e.g., Z-score)
 - Mais en utilisant le même système de dépistage (e.g., bibliothèques numériques) (condition C), la simple fusion par le score offre une bonne performance
- Avec de l'apprentissage, voir CMU à CLEF 2005
- Berkeley à CLEF 2003
 - RI multilingue, 8 langues
QT: 0,3317 DT (EN en pivot): 0,3401
le mix DT & QT (et fusion): 0,3733
- Avec le mix QT et DT on obtient les meilleures performances (voir CLEF 2003 multilingual track)

EARIA'06 - 72 -

RI multilingue (CLEF-2003)

Multilingual-8; Recall-Precision Graph

Top performing experiments by the best five groups; TD topic fields; Automatic



RIA'06 - 73 -

Conclusion

- Les stratégies d'appariement sont nettement indépendantes de la langue
- Recherche monolingue
 - Adaptation simple pour des langues proches de l'EN (les familles romanes et germaniques)
 - Est-ce la même chose pour la famille slave ?
 - Les mots composés sont importants en DE
 - Des analyseurs morphologiques plus poussés s'avèrent clairement utile dans certains cas (FI)
 - la segmentation est un problème (ZH, JA)
 - pas de conclusion définitive pour KR, HU
 - quelques collections test ne sont pas de bonne qualité (AR à TREC 2001, RU à CLEF 2004)

Conclusion

- Recherche bilingue / multilingue
 - de nombreux outils sont disponibles pour des paires de langues (essentiellement avec EN)
 - plus de problème avec des langues moins importantes et moins utilisées sur le Web
 - la performance d'une recherche bilingue peut être proche en qualité d'une recherche monolingue
 - la fusion n'est pas encore résolue (voir CMU à CLEF 2005)
 - on ignore un grand nombre de langue (e.g., l'Afrique)

EARIA'06 - 75 -

Et ...

Thank you

Gracias

Grazie

Dask u

Kiitos

Danke

Merci

Task

謝謝

EARIA'06 - 76 -

Références

- Abdou, S., Savoy, J. Statistical and comparative evaluation of various indexing and search models. Proceedings AIRS-2006
- Amati, G., van Rijsbergen, C.J. (2002) Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM - Transactions on Information Systems, 20, 357-389.
- Brown, P., Della Pietra, S., Della Pietra, V., Lafferty, J., Mercer, R. (1993) The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2), 263-311.
- Braschler, M., Ripplinger, B. (2004) How effective is stemming and compounding for German text retrieval? IR Journal, 7, 291-316.
- Braschler, M., Peters, C. (2004) Cross-language evaluation forum: Objectives, results, achievements. Information Retrieval, 7(1-2), 7-31.
- Braschler, M. (2004) Combination approaches for multilingual text retrieval. Information Retrieval, 7(1-2), 183-204.
- Gao, J., Nie, J.-Y. (2006) A study of statistical models for query translation: Finding a good unit of translation. ACM-SIGIR'2006. Seattle (WA), 194-201.
- Gale, W.A., Church, K.W. (1993) A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(1), 75-102.
- Grefenstette, G. (Ed) (1998) Cross-language information retrieval. Kluwer.
- Harman, D. (1991) How effective is suffixing? Journal of the American Society for Information Science, 42, 7-15.

Références

- Harman, D. (1991) How effective is suffixing? Journal of the American Society for Information Science, 42, 7-15.
- Harman, D.K. (2005) Beyond English. In "TREC experiment and evaluation in information retrieval", E.M. Voorhees, D.K. Harman (Eds), The MIT Press.
- Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A., Järvelin, K. (2004) Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000–2002. Information Retrieval, 7 (1-2), 99-119.
- Hiemstra, D. (2000) Using language models for information retrieval. CTIT Ph.D. thesis.
- Kraaij, W. (2004) Variations on language modeling for information retrieval. CTIT Ph.D. thesis.
- Krovetz, R. (1993) Viewing morphology as an inference process. ACM-SIGIR'93. Pittsburgh (PA), 191-202.
- Le Calvé A., Savoy J. (2000) Database merging strategy based on logistic regression. Information Processing & Management, 36(3), 341-359
- McNamee, P., Mayfield, J. (2004) Character n-gram tokenization for European language text retrieval. IR Journal, 7(1-2), 73-97.
- Nie, J.Y., Simard, M., Isabelle, P., Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. ACM-SIGIR'99, 74-81.

Références

- Porter, M.F. (1980) An Algorithm for suffix stripping. *Program*, 14, 130-137.
- Savoy, J. (1993) Stemming of French words based on grammatical category. *Journal of the American Society for Information Science*, 44, 1-9.
- Savoy J. (2004) Combining multiple strategies for effective cross-language retrieval. *IR Journal*, 7(1-2), 121-148.
- Savoy J. (2005) Comparative study of monolingual and multilingual search models for use with Asian languages. *ACM -Transaction on Asian Language Information Processing*, 4(2), 163-189.
- Savoy J. (2006) Light stemming approaches for the French, Portuguese, German and Hungarian languages. *ACM-SIAC*, 1031-1035.
- Sproat, R. (1992) *Morphology and computation*. The MIT Press.
- Xu, J., Croft, B. (1998) Corpus-based stemming using cooccurrence of word variants. *ACM -Transactions on Information Systems*, 16, 61-81.
- Xu, J., Weischedel, R., Nguen, C. (2001) Evaluating a probabilistic model for crosslingual retrieval. *ACM –SIGIR-2001*, New Orleans, 105-110.
- Zhang, Y., Vines, P., Zobel, J. (2005) Chinese OOV translation and post-translation query expansion in Chinese-English cross-lingual information retrieval. *ACM - Transactions on Asian Language Information Processing*, 4 (2), 57-77.

Ressources du Web

Journal officiel de UE: eur-lex.europa.eu

Organisation des nations unies: www.un.org

EuroWordNet www.illc.uva.nl/EuroWordNet/

Campagnes d'évaluation: CLEF, NTCIR, TREC

Trésor de la langue française: atilf.atilf.fr/tlf.htm

Campagne d'évaluation (CLEF 2005)

	FR	PT	BG	HU
Taille en MB	487 MB	564 MB	213 MB	105 MB
Documents	177 452	210 734	69 195	49 530
mots distincts / document	178	213	134	142
Nombre de requêtes	50	50	49	50
Nombre de réponses / requête	50,74	58,08	15,88	18,78

EARIA'06 - 81 -

RI monolingue (raciniseur)

- Enracineur
 - Avec des exceptions (dans toutes les langues)
box → boxes, child → children
one walkman → ? (walkmen / walkmans)
et d'autres problèmes: "The data is/are ...",
en FR: amour, délice, orgue
 - Approches suggérées (inflexion + dérivation)
Lovins (1968) ? 260 règles
Porter (1980) ? 60 règles
variante: S-stemmer [Harman 1991]: 3 règles
 - Evaluation et problème en EN [Harman 1991]

EARIA'06 - 82 -

RI monolingue (raciniseur)

- Basé sur la grammaire: Règles (ad hoc)
 - concerne les suffixes (pré-retraite et retraite)
 - contrôle via des contraintes quantitatives
 - contrôle via des contraintes qualitatives
 - règles de réécriture
- La mesure de performance est souvent une mesure moyenne (MAP) et la solution générale n'est pas forcément la meilleure pour une application dans un domaine spécifique
- On rencontre des erreurs (on obtient pas la racine attendue) par suppression trop forte ou trop faible
 - “organization ” → “organ”
 - “féministe ” → “fémin”

RI monolingue (raciniseur)

- Exemples
 - IF (" *-ing ") then éliminer -ing
e.g., "king" ? "k", "running" → "runn"
 - IF (" *-ize ") then éliminer -ize
e.g., "seize" → "se"
- mais on corrige ces cas par :
 - IF ((" *-ing ") & (length>3)) then éliminer -ing
 - IF ((" *-ize ") & (!final(-e))) then éliminer -ize
 - IF (suffix & control) ? remplacer ...
"runn" → "run"

Campagne d'évaluation

Les requêtes sont disponibles en plusieurs langues
(CLEF 2005)

- EN: Nestlé Brands
FR: Les Produits Nestlé
PT: Marcas da Nestlé
HU: Nestlé márkák
BG: Продуктите на Нестле
- EN: Italian paintings
FR: Les Peintures Italiennes
PT: Pinturas italianas
HU: Olasz (itáliai) festmények
BG: Италиански картини

EARIA'06 - 85 -

Campagne d'évaluation

- Requêtes avec une couverture relativement large
 - « Embargo sur l'Iraq »
 - « Référendum norvégien à propos de l'UE »
 - « Les victoires d'Alberto Tomba »
 - « Golden Globes 1994 »
- Requête plus spécifique d'un pays / région
 - « L'activité agricole dans le Delta de l'Ebre »
 - « Initiative suisse pour les Alpes »

EARIA'06 - 86 -